浅析基于商业智能的新闻采编业务流程数据分析挖掘

摘　要：随着大数据时代的到来，越来越多的企业采用商业智能的相关技术，从生产、销售等环节的数据中“淘金”，为企业决策层提供辅助决策。本文将商业智能关键技术应用于新闻采编业务，针对新闻生产业务的全流程，通过对全流程状态分析、新闻生产力、传播影响力等主题的数据分析挖掘，展示新闻采编发的流转过程，展示采编部门和人员的生产效率，展示稿件的传播影响力。文中介绍了商业智能的工作原理，对新闻采编业务流程数据分析挖掘方法进行了深入研究。

关键词：商业智能；数据分析；数据挖掘
中图分类号：G210.7　　文献标识码：A
文章编号：1671-0134（2017）12-118-03　　DOI：10.19483/j.cnki.11-4653/n.2017.03.030

■文 / 陈辛夷　康　洁

引言

在大数据时代，数据的价值越来越受到各行业的重视。企业内积累的大量业务流程数据迫切需要人们从中“淘金”。商业智能是能满足企业这一迫切需求的有力工具，能将海量数据转化为知识，有助于从以往数据中发现业务趋势，为企业决策层提供辅助决策。Gartner 调查显示，2012 年和 2013 年，BI 已上升到全球 CIO 优先考虑的十大技术的首位。

当今传统媒体转型面临严峻形势，而大数据将是媒体转型的有力武器。在新媒体时代，用户需要及时、准确、个性化的新闻服务。数据挖掘可以帮助传统媒体充分发挥人才资源优势，提升品牌竞争力和用户黏性。在新闻采编业务中，存在大量流程数据，在以往的采编系统中这些数据并未受到重视，而借助商业智能的相关技术对这些业务流程数据进行分析挖掘，有助于提高采编全流程业务管理信息化水平，掌握报道进展情况、人员工作效率、稿件落地情况和传播效果等。

1. 商业智能的定义

商业智能又名商务智能（Business Intelligence，BI）。商业智能对数据进行收集、管理，提供一系列技术和方法对企业的各类数据进行分析。商业智能可以帮助企业的领导层从宏观上掌握企业的运转情况，洞察潜在行业的机会，辅助他们进行决策。

2. 研究商业智能的意义

商业智能帮助企业迅速发现问题，提示企业管理者加以解决。具体到新闻采编行业来说，商业智能贴近媒体管理者的迫切诉求。通过对新闻传播影响力的分析挖掘，使管理者可以清楚掌握新闻的传播效果和影响力。

商业智能为新闻创造价值，帮助传统媒体实现以“终端用户为中心”的转型升级，通过对用户行为的分析挖掘，可以对用户群体按照性别、年龄、职业、地域等因素进行分类或聚类，把用户进行群体细分，针对不同用户推荐感兴趣的新闻内容，使媒体更懂用户。

帮助在新闻生产的每个环节控制成本，通过新闻生产力的分析挖掘，展示各采编部门和人员的生产效率，为采编人员和部门考核提供依据。运用商业智能的方法，可以提高决策的水平，对业务流程进行改进，最终提高管理的效率。

及时性是新闻的基础，通过对互联网海量数据的挖掘可以发现潜在的新闻热点。比如：网络媒体和新媒体中大量用户的阅读和评论数据可以辅助采编人员发现新闻热点。

3. 商业智能关键技术

3.1 OLAP

即联机分析，提供多维数据管理环境，使企业的数据分析人员能从多个维度对商业问题进行建模和分析。

3.2 数据分析

使用适当的统计分析方法对数据进行分析，提取出有价值的信息。

3.3 数据挖掘

数据挖掘就是从大量数据中挖掘出隐含的、未知的、有价值的关联和模式，建立可用于决策的模型，提供分析风险、进行预测的功能。

4. 商业智能体系结构

首先将分散在企业各系统中的数据，包括关系型数据也包括非关系型数据进行汇总，通过数据抽取（Extract）、转换（Transform）、清洗（Cleaning）、装载（Load），最终按照预先定义好的数据模型，将数据加载到数据仓库中，这一过程简称 ETL。

通过对企业数据需求的分析，建立企业数据仓库的逻辑模型和物理模型，将企业各类数据按照分析主题进行组织和归类。

在数据仓库的基础上提供多种软件工具供终端用户查询和生成报告，包括 OLAP 工具、数据挖掘软件、报表工具等。

5. 在新闻业务中的应用

5.1 数据源

数据仓库中数据的采集需要从各种业务应用系统和管理信息系统中获取，如稿件建采系统、编辑系统、供稿系统、OA 系统等，按照统一的数据标准存放在数据仓库中。

本文将采编业务系统数据划分为静态信息数据、动态信息数据两大类。

静态信息数据是指相对稳定的信息，主要指采编部门、采编人员、发稿线路等静态属性信息数据。

动态信息数据收集在采编业务系统中不断变化的流程数据，包括采、编、签、改、发、供、馈等环节。如何对新闻业务数据，特别是用户行为数据构建数据模型，分析稿件流转过程，将是本文着重介绍的内容。

5.2 分析目标

通过采集稿件、流程、人员和质量数据，进行采编业务全流程的管理，掌握报道进展情况、人员工作效率、稿件落地情况、传播影响效果等。从全流程状态、传播影响力、新闻生产力等主题进行数据分析挖掘，呈现新闻生产业务运行状况。

新闻生产力分析：分析呈现采编部门、采编人员等在一段时间内的工作效率。

传播影响力：分析呈现稿件的落地情况和传播影响效果。

全流程状态分析：分析稿件在各采编环节的流转情况。

5.3 数据建模

数据建模主要用到的是维度模型。一个度量往往和多个维度相关，维度模型表达了数据之间的关联关系。比如：想要了解 2016 年 1 月份在新媒体线路的中文稿件发稿情况，这个发稿量数据与线路、时间、语种三个维度相关。维度建模是从多个角度和层次反映数据之间的联系，从多个维度对数据进行重组，为决策提供数据的多维视图。

维度模型有两种不同性质的表：事实表和维度表。

通常采用星型或雪花模型把事实表和维度表融合在一起，中间是事实表，周围是维度表。

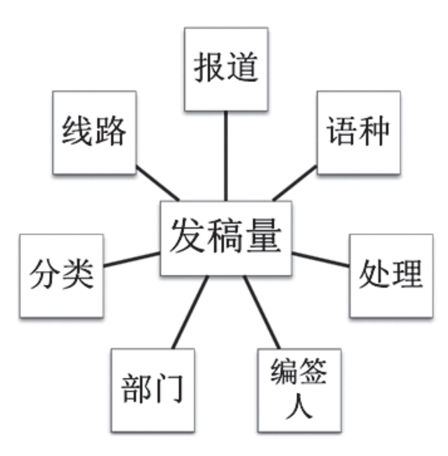


图 1　新闻发稿情况雪花模型示意

事实表存放的是业务性能的度量值。 一个事实代表一个业务度量值，如：发稿量。

维度表提供观察度量值即事实的角度，如：线路、时间。

维度表的属性列（如：语种包含中、英、法、意、日、阿、俄）是用户使用数据的约束条件，同时也是数据分析时的切割工具，因此维度表的质量与深度直接影响整个数据仓库的性能。

对于稿件的业务处理流程，本文采用处理维度进行描述，属性列对应采编业务流程中的采稿、入库、建稿、建新稿、编辑、签发等环节。

在数据仓库中可以根据需要，建立多个应用主题，本文建立了新闻生产力分析主题、传播影响力分析主题和全流程状态分析主题。

5.4 关键指标体系

5.4.1 新闻生产力

在新闻生产力评估中可以采用生产率、人均生产稿件数量、投入人员占比等作为考核部门新闻生产力的指标，指标值可根据时间汇总到年、季、月、双周、周、日，可通过以下维度查看稿件数量的详细情况：媒体类型、新闻分类、供稿类别、稿件处理流程。

表 1　新闻生产力指标

指标	指标说明
稿件数量	部门（发稿部门、签发部门）单位时间内处理的稿件数量
总消耗时间	部门（采稿、签发部门）单位时间内处理稿件所消耗的时间
生产率	总消耗时间 / 稿件数量
记者人数量	发稿部门处理稿件所投入的人员数量
编辑人数量	编辑部门处理稿件所投入的人员数量
签发人数量	签发部门处理稿件所投入的人员数量
人均生产稿件数量	稿件数量 / 部门总人数
投入人数占比	处理稿件投入人员总数 / 部门总人数

5.4.2 传播影响力

在传播影响力评估中，在本文中采用传统媒体影响力指数、网络媒体影响力指数、国内媒体影响力指数、海外媒体影响力指数、海外社交媒体影响力指数、全网影响力指数为主要的指标。其中全网影响力指数为其余五个指数的加权计算结果。

网络媒体传播影响力指数如下表，指标值可根据时间汇总到年、季、月、双周、周，可查看指标在不同媒体上的详细情况。

表 2　网络媒体传播影响力指标

指标	指标说明
采用量	单位时间内，单一稿件在网络媒体上的采用量
评论量	单位时间内，单一稿件在网络媒体上的评论量
评论同向	单位时间内，单一稿件在网络媒体上的评论信息中正面评论占比
评论异向	单位时间内，单一稿件在网络媒体上的评论信息中负面评论占比
曝光度	单位时间内，单一稿件在网络媒体上的受众的总量
转载（转引）	单位时间内，单一稿件在网络媒体上的转载量
转载（转引）深度	单位时间内，单一稿件在网络媒体上的转载深度

5.4.3 全流程状态分析

全流程状态分析可实时监测各指标的变化情况，可通过以下维度查看指标的详细情况：稿件处理（采稿、入库、建稿、建新稿、编辑、签发等）、稿件媒体类型、稿件供稿类别。

表 3　稿件组全流程状态分析指标

指标	指标说明
稿件数	前一稿件处理节点至完成当前处理节点所完成的稿件数量
综合处理耗时	前一稿件处理节点至完成当前处理节点所完成的稿件组所消耗的时间
综合累计耗时	稿件组从发稿时点至现在所消耗的时间累计。即：综合处理耗时的累加
平均综合处理耗时	综合处理耗时 / 稿件数
平均综合累计耗时	综合累计耗时 / 稿件数
阶段综合百分比	每个业务环节（采、编、发、供、馈）设定标准工作时间定额。稿件组在各业务环节内完成时间百分比
全流程综合百分比	稿件组在整个业务流程的完成时间百分比

5.5 业务流程数据挖掘算法

在“以用户为中心”的思想指导下，充分利用关联规则、分类、聚类等数据挖掘技术，对日常新闻业务数据进行挖掘。本文采用以下方法对新闻业务用户行为数据、全流程状态数据等进行分析。

5.5.1 关联规则和序列模式

关联规则用于分析用户数据，发现用户行为模式。关联规则描述数据项之间存在的关联关系，即根据一个事务中某些项的出现推导出另一些项在同一事务中也出现。Apriori 算法是关联规则的经典算法。关联规则最初针对购物篮分析问题提出，即分析消费者经常同时购买哪几种商品。在新闻业务中，关联规则挖掘可以找出新闻采编业务人员个人特征与稿件之间的关联性；根据业务人员的关注点推荐相关稿件，将相同性质的报道任务分配给适当的记者或编辑。

5.5.2 时间序列分析

时间序列分析根据固定时间间隔来记录事件结果。新闻业务系统每天固定时段处理稿件数变化，每月处理稿件数，每季度总的发稿量等就是时间序列的案例。

分析时间序列数据，可以借助一些可视化的手段，如：柱状图、折线图，从而观察出某些现象特征及行为，通常时间序列有四种主要的变化：

长期或趋势变化。用于反映长期变化的总体方向，体现为趋势线。

循环运动。体现为沿着趋势线或者趋势曲线长时间的摆动，包括周期性和非周期性的摆动。

季节性移动或季节性变化。反映每年都重复出现的事件，体现为在连续几年的同期重复出现相同或相似的模式。

非规律或随机变化。由于偶然或随机事件引起的变化。

数据挖掘技术应用于新闻业务流程管理对数据的归纳、分析和处理精细化有重要帮助。通过获取与分析用户行为模式，分析以往采编流程数据，全面掌握采编业务的运作状态，了解采编人员的特点，实现服务个性化、智能化。

6. 结束语

在传统媒体战略转型的迫切形势下，需要依靠技术创新提升核心竞争力和传播影响力。大数据是内容、渠道、服务的核心支点，是传统媒体转型的有力推手。本文探讨了在商业智能的通用框架下，数据分析挖掘技术在新闻采编业务流程数据上的应用。通过对新闻生产力、传播影响力、全流程状态的分析挖掘，使用先进的方法和工具，梳理采编业务流程，识别行为数据产生点和管理控制点并进行指标体系设计，帮助决策者把握业务发展方向。随着大数据时代的发展，商业智能相关技术的应用将助力媒体融合，为传统媒体战略转型提供有力支持。媒

参考文献

[1] 张良均，陈俊德等 . 数据挖掘实用案例分析 [M]. 北京：机械工业出版社，2013（7）： 18–30.

[2] 陈哲 . 数据分析企业的贤内助 [M]. 北京：机械工业出版社，2015（5）：1–27.

[3] Ralph Kimball，Margy Ross. 数据仓库工具箱（第三版）[M]. 北京：清华大学出版社， 2015（1）：5–11.

（作者单位：新华社技术局）